JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

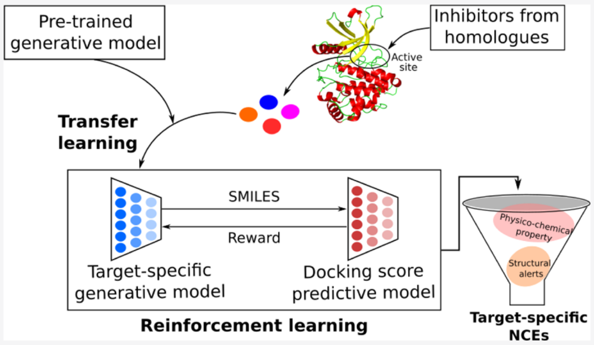
Accelerating *De Novo* Drug Design against Novel Proteins Using Deep Learning

Sowmya Ramaswamy Krishnan,[†] Navneet Bung,[†] Gopalakrishnan Bulusu,* and Arijit Roy*

Cite This: *J. Chem. Inf. Model.* 2021, 61, 621−630 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: In the world plagued by the emergence of new diseases, it is essential that we accelerate the drug design process to develop new therapeutics against them. In recent years, deep learning-based methods have shown some success in ligand-based drug design. Yet, these methods face the problem of data scarcity while designing drugs against a novel target. In this work, the potential of deep learning and molecular modeling approaches was leveraged to develop a drug design pipeline, which can be useful for cases where there is limited or no availability of target-specific ligand datasets. Inhibitors of the homologues of the target protein were screened at the active site of the target protein to create an initial target-specific dataset. Transfer learning was used to learn the features of the target-specific dataset. A deep predictive model was utilized to predict the docking scores of newly designed molecules. Both these models were combined using reinforcement learning to design new chemical entities with an optimized docking score. The pipeline was validated by designing inhibitors against the human JAK2 protein, where none of the existing JAK2 inhibitors were used for training. The ability of the method to reproduce existing molecules from the validation dataset and design molecules with better binding energy demonstrates the potential of the proposed approach.


Pre-trained generative model
Inhibitors from homologues
Active site
Transfer learning
SMILES
Reward
Target-specific generative model
Docking score predictive model
Reinforcement learning
Physico-chemical property
Structural alerts
Target-specific NCEs

■ INTRODUCTION

Outbreak of emerging diseases poses a severe threat to the human population and adversely affects the economic situation of several countries. It is extremely important to come up with measures to curb the spread of such diseases. Although developing a drug is one of the most promising ways for the treatment of a disease, it is also a time-, cost- and resource-intensive process.[1] The long multistep process involves hit identification, lead optimization, and preclinical and clinical trials before the drug reaches the market.[1] The hit identification and lead optimization processes can extend to 2 years or more. Accelerating these processes will have a larger influence on the time and cost of the downstream phases of the drug discovery process.

Small molecules are primarily designed to modulate the function of a specific target protein in the biological system. Traditional drug design methods identify molecules specific to a target protein of interest by screening libraries of compounds available in both public and commercial repositories or by *de novo* generation of molecules using fragments and pharmacophore models.[2] While the screening process is very time-consuming, only a part of the chemical space, approximately a billion small molecules, has been explored using traditional approaches.[3] However, the actual chemical space is estimated to be ~10^{63} molecules or more.[3] Deep learning approaches can bridge this gap and design diversified new chemical entities with desired drug-like properties. Further, the challenge is to develop a method which can design suitable small molecules specific to any target protein of interest. Developing such a method can rapidly design drugs for target proteins of emerging diseases, where no previous target-specific datasets are available.

Recent developments in the field of artificial intelligence (AI) and big data have shown the potential to radically transform the accuracy and reliability of computational models in several fields of health care,[4−6] including drug discovery.[7] The simplified molecular input line entry system (SMILES) representation[8−11] or the molecular graph representation[12] is commonly used for training the deep neural network models to learn feature representations. Drug discovery also requires control over multiple structural and physicochemical parameters.[13] Although earlier studies were focused on the generation of libraries for virtual screening,[8] the introduction of reinforcement learning for property optimization has helped in biasing the models to generate compounds with the properties of interest.[9−11,13] Further, the efficiency of the models to generate chemically valid molecules can be significantly improved by using memory-

Received: September 10, 2020
Published: January 25, 2021

https://dx.doi.org/10.1021/acs.jcim.0c01060

augmented neural networks.[10,11,14] Although there have been several advancements in the application of AI-based methods, the availability of data for protein-specific drug discovery still remains a challenge.

In this study, we have developed a *de novo* drug design pipeline which can be used against novel target proteins whose 3D structure is available/modeled and the active site is characterized. A generative model was initially trained to learn the grammar of known drug-like molecules. For cases where there is no available target protein-specific ligand dataset, the known inhibitors of the same family of proteins as the target protein were considered. Next, molecular modeling technique was applied to compile the dataset specific to the target protein. Transfer learning was used to learn the features of the compiled dataset. Finally, reinforcement learning was used to combine the generative and predictive models, which optimizes the scoring function to produce small molecules that are predicted to bind better to the target protein with desirable drug-like properties. As a proof of concept, the *de novo* drug design pipeline was used to design inhibitors against the human Janus kinase 2 (JAK2) protein.

■ MATERIALS AND METHODS

The proposed *de novo* drug discovery pipeline can be divided into the following components: (1) dataset curation, preprocessing, and training the generative model; (2) generation of target-specific ligand dataset and transfer learning; (3) training the predictive model; (4) reinforcement learning for property-optimized molecule generation; (5) filtering the generated molecules through physicochemical properties; (6) application of rule-based filters to remove molecules with undesirable groups; and (7) validation of the filtered molecules through molecular modeling. Each step is discussed in detail below.

Dataset Curation, Preprocessing, and Training the Generative Model. The first step of the pipeline is to train a generative model such that it learns the grammar of small molecules and designs novel and valid small molecules. The datasets for training the generative model and for the case study were obtained from the ChEMBL database.[15] The molecules were represented in the SMILES format[16] to leverage the effectiveness of recurrent neural networks in handling sequential data through the existing natural language-processing algorithms. The problem of learning the SMILES grammar is cast as a Seq2Seq problem (machine translation), and the generative model tries to emulate this problem for the SMILES dataset with stack-augmented recurrent neural networks.[14] The SMILES dataset was preprocessed by applying sequential filters to remove stereochemistry, salts, and molecules with undesirable atoms or groups.[10] The SMILES strings obtained were canonicalized, and duplicates were removed. Only molecules with length ≤ 100 were collected. The RDKit library[17] in Python was used for dataset preprocessing.

The generative model was trained on a dataset of ∼1.6 million SMILES strings from the ChEMBL database[15] (step a, Figure 1). The deep neural network architecture of the generative model uses gated recurrent unit (GRU) as the internal memory,[18] augmented with a stack acting as the dynamic external memory.[14] The use of stack-augmented memory[14] enabled the generation of chemically valid SMILES with high accuracy, as also observed in previous studies.[10,11] The trained generative model was used to generate 100,000 compounds in 10 batches of 10,000 compounds each. This model is henceforth referred to as the pretrained generative model. The pretrained generative model had a high accuracy of 96.6%, defined as the mean percentage of chemically valid molecules present in all the sampled batches. The other metrics of the pretrained generative model are provided in Supporting Information 1—Section S1.

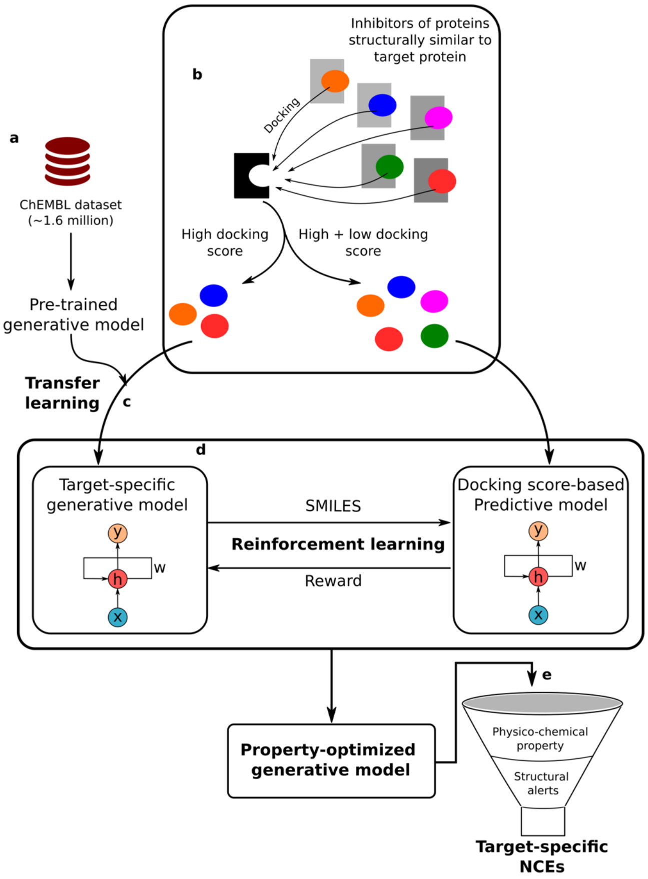


Figure 1. *De novo* small-molecule design pipeline: (a) training the generative model on the ChEMBL database to learn the grammar of small molecules; (b) dataset of inhibitors was curated from small molecules that modulate the activity of the same family of proteins; (c) transfer learning with the curated dataset helps focus on a particular region of the chemical space specific to the target protein of interest; (d) reinforcement learning enables generation of novel small molecules with optimized docking scores. The *x*, *h*, *y*, and *w* correspond to the input, hidden state, output, and weights, respectively; (e) physicochemical properties and structural alerts (rule-based filters) were used to filter drug-like molecules specific to the target protein of interest.

Generation of Target-Specific Ligand Dataset and Transfer Learning. The drug design pipeline aims to discover novel small molecules against a specific target protein. In most cases, there is limited or no knowledge about the small molecules that can bind to the target protein of interest. In this study, an initial target-specific small-molecule dataset was curated, considering molecules known to inhibit a similar protein family to the target protein (step b, Figure 1). In order to identify the proteins belonging to the same protein family as the target protein, the Basic Local Alignment Search Tool (BLASTp)[19] was utilized, with the sequence of the target protein as the search key. UniProtKB/Swiss-Prot database was used as the search database, and the other parameters for the BLAST were kept at their default values.[20]

The curated dataset of small molecules from similar proteins was subjected to extensive preprocessing to canonicalize and remove redundancy. The pChEMBL score of the small molecules based on the reported IC_{50} values was also used to filter the dataset, such that the molecules with high bioactivity (pChEMBL score ≥ 6.0) were chosen for training the generative model. The target selectivity of this dataset was further enhanced by docking these molecules in the active site of the protein of interest using AutoDock Vina.[21] The molecules with high docking scores were used to re-train the generative model to capture the molecular features specific to a receptor of interest through transfer learning (step c, Figure 1).

During transfer learning, the weights of the pretrained, unbiased generative model (prior) were loaded, and all the layers of the deep neural network were frozen, except for the last dense layer.[22] This allows the gradient calculation and back-propagation steps to alter the weights of only the final dense layer, thereby preventing the catastrophic "forgetting" of features learned during model training. The model was trained until the inferred molecules show an observable shift in similarity with respect to the training dataset, quantified using the Tanimoto coefficient.[23]

Training the Predictive Model. The predictive model learns a mapping between the small molecules (represented as SMILES strings) and their corresponding experimentally determined property values, such as bioactivity.[10] As this work focuses on cases where no target-specific dataset is available, both the high and low docking scores of ligands from the previous step were used to train the predictive model. The predictive model uses GRU as the internal memory,[18] followed by three dense layers (see Supporting Information 1). The model was trained using a mini-batch gradient descent with the Adam optimizer.[24] All the architecture and hyperparameter information for the generative and predictive models are provided in Supporting Information 1 (Table S1).

Reinforcement Learning: Property-Optimized Molecule Design. The generative model obtained after transfer learning was combined with the predictive model to bias the generative model toward the property of interest using reinforcement learning[9−11] (step d, Figure 1). During training, the generative model acts as the agent, and the predictive model acts as the critic. The agent learns a prior policy during training, which is the probability distribution over the different symbols at each position of the SMILES string (trajectory/episode). The objective of the policy gradient method is to refine or optimize this prior policy so that the reward obtained is maximum.[25]

The state space, S, is the set of all possible strings based on the SMILES vocabulary, and the action space A is the set of all probable characters for the next position given the character of the current position. The task of generating the SMILES string is considered to be episodic in the sense that the reward can be computed only after a complete SMILES string has been sampled. The reward is a value computed using a reward function defined in terms of the property values predicted by the critic. The reward function for training the agent is

$$r(x) = \exp\left(\frac{-x}{3.0}\right) \tag{1}$$

where x is the predicted docking score from the predictive model (critic).[10] This reward function helps in the optimization of the virtual screening scores of the generated molecules. The canonical policy gradient algorithm tends to introduce a catastrophic "forgetting" behavior in deep neural networks during the course of the joint training process.[26] By anchoring the new policy to the prior policy of the agent, reinforcement learning methods have overcome this issue.[9,27]

During the policy gradient training, two copies of the generative model were considered. The weights of the first copy of the model, termed as the prior, were kept unchanged. The weights of the second copy of the model, termed as the agent, were varied with regularization, so that the new policies were highly similar to the prior policy in terms of the learned grammar of small molecules while also ensuring that the generated molecules attain the property sweet spot.[9]

For an action space A, the model likelihood for sampling a given SMILES string is given by the product of the action probabilities. Let $f(A)$ be the reward function based on the predictive model. The agent was anchored to the prior through an augmented likelihood, calculated as below 2.

$$\log P(A)_{\mathrm{U}} = \log P(A)_{\mathrm{prior}} + \sigma f(A) \tag{2}$$

where $P(A)_{\mathrm{prior}}$ denotes the model likelihood calculated using the prior policy, and σ denotes the weight factor for the reward function. It helps the agent find a balanced policy between the prior and agent policies. Thus, the long-term return can be modified as follows 3.

$$r(s_{\mathrm{T}}) = -[\log P(A)_{\mathrm{U}} - \log P(A)_{\mathrm{A}}]^2 \tag{3}$$

where $P(A)_{\mathrm{A}}$ denotes the model likelihood, calculated using the agent policy. This episodic reward can be maximized by casting it into a minimization problem, resulting in the following loss function or objective for training the agent 4.

$$J(\theta) = -r(s_{\mathrm{T}}) \tag{4}$$

The regularized policy gradient method was trained using a mini-batch gradient descent with AMSGrad optimizer.[28] All the architecture and hyperparameter information for transfer learning and reinforcement learning are provided in Supporting Information 1 (Table S1).

Filtering the Generated Molecules through Physicochemical Property Filters. The generated small molecules were canonicalized to remove duplicates and molecules identical to the transfer learning training dataset. The small molecules were further subjected to stringent physicochemical property filters, whose thresholds were decided based on the physicochemical property distributions of known target-specific inhibitors. The various filters applied include: synthetic accessibility score,[29] octanol−water partition coefficient, and molecular weight (step e, Figure 1). The generated small molecules which passed all the physicochemical property filters were taken for the subsequent steps of filtration and analysis.

Application of Rule-Based Filters to Remove Molecules with Undesirable Groups. The set of molecules obtained after the application of physicochemical property filters was subject to screening using the Pan Assay Interference Compounds (PAINS) filter,[30] BRENK filter,[31] National Institute of Health (NIH) filter,[32] and ZINC filter (step e, Figure 1). These filters employ empirically observed rules to avoid toxic and synthetically infeasible subgroups in the small molecules. RDKit[17] was used to apply all the four filters on the filtered set of small molecules, and those molecules flagged by at least 2 filters were removed.

Validation of the Molecules Through Molecular Modeling. To study the effect of solvents and also to compute binding energies, molecular dynamics simulations were performed using GROMACS 2016[33] and CHARMM36 force-

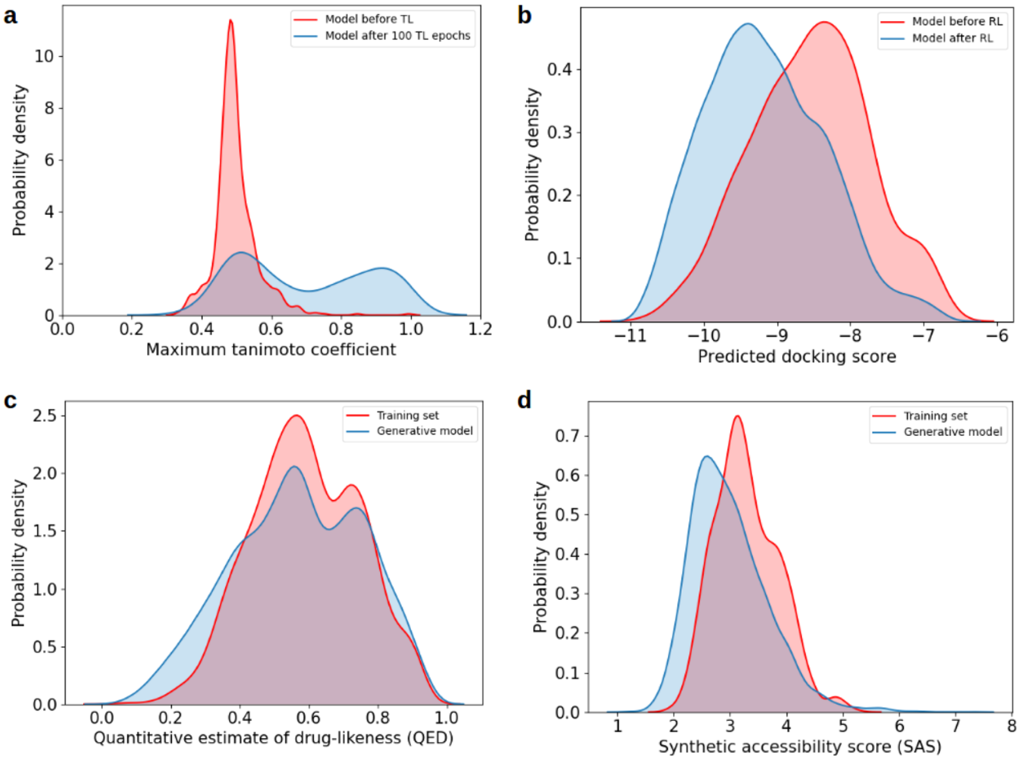


Figure 2. Shift in property distributions during transfer learning and reinforcement learning: (a) distribution of the maximum Tanimoto coefficient between the molecules generated by the model and the transfer learning training dataset before and after transfer learning; (b) distribution of the predicted docking scores before and after 70 epochs of reinforcement learning, indicating the amount of shift as a result of optimization after reinforcement learning. (c) Distribution of the quantitative estimate of drug-likeness scores and (d) distribution of the synthetic accessibility scores were also monitored to check the quality of the molecules generated after reinforcement learning.

field.[34] The parameters for the small molecules were obtained using CgenFF.[35] The system was solvated in a cubic box of TIP3P model[36] water molecules. The system was then energy-minimized using the steepest descent, followed by NVT and NPT position-restrained equilibration for 200 ps and 1 ns, respectively.[37] After equilibration, a production run of 5 ns was performed on the docked complexes using the NPT ensemble. Binding energies for the trajectories obtained from MD simulations for each of the complexes were calculated using the g_mmpbsa[38] module of GROMACS. For each case, two separate calculations were performed and then averaged to find the final binding energy.

■ EXPERIMENTAL SETUP

To validate the proposed pipeline, human JAK2 protein is used as the target protein of interest. JAK2 is a tyrosine kinase protein involved in various essential cellular processes including cell growth, development, differentiation, and histone modifications.[39] It plays a major role in regulating both the innate and adaptive immune systems.[40] The dysfunction of JAK2 has been implicated in multiple conditions such as myelofibrosis and thrombocythemia.[41] For the current case study, all the existing JAK2 inhibitors were collected but used only for validating the pipeline.

Identification of the Homologues of the Target Protein. A ligand dataset specific to the target protein is necessary for training the transfer learning model. As it was assumed that there was no prior knowledge about the inhibitors of JAK2, a set of inhibitors of proteins that belong to the JAK2 protein family was identified. First, the sequence of JAK2 protein from the UniProt database[42] (UniProt ID: O60674) was used for BLASTp search[19] against all human proteins, with a query coverage of 90% and an *E*-value below 0.01.[20] It was identified that the kinase domains of three other human proteins, namely, JAK1, JAK3, and TYK2, are highly similar in sequence to the kinase domain of the human JAK2 protein with some differences (Supporting Information 1—Section S2). A comparison of the active sites of the four proteins of the human Janus kinase protein family reveals that the kinase domains have a conserved active site structure (Supporting Information 1—Figure S2).

Curation of the Target-Specific Training and Validation Datasets. The ligands known to inhibit JAK1, JAK3, and TYK2 proteins were chosen to construct the target-specific training dataset. All the datasets with their experimentally determined IC_{50} values were collected from the ChEMBL database.[15] The IC_{50} values were converted to log scale to obtain the pChEMBL score. For a molecule found to be assayed against multiple JAK family proteins, the protein against which the molecule had maximum pChEMBL score was selected as the target protein for the molecule. After preprocessing, canonicalization, and removal of redundant molecules among JAK1, JAK3, and TYK2 inhibitors, a final dataset of 4167 molecules was obtained, of which 3711 molecules were found to have a pChEMBL score ≥ 6.0. A validation dataset of 1103 unique JAK2 inhibitors from the ChEMBL database was also curated.

To ensure the specificity of the 3711 inhibitors, the dataset was screened by docking in the active site of human JAK2 protein using AutoDock Vina.[21] Only 3681 molecules with a virtual screening score ≤ -7.0 were utilized for training the generative model using transfer learning. The dataset of 4167 molecules with both high and low docking scores was used for a docking score-based predictive model.

■ RESULTS

Designing Small Molecules with Optimized Docking Score Specific to Janus Kinase 2 (JAK2) Protein. From the curated dataset of 3681 molecules (step c, Figure 1) transfer learning was performed for 100 epochs until the distribution of the Tanimoto coefficient[23] between the inferred molecules and the training dataset shows no further improvement (Figure 2a). While the target-specific generative model designed new molecules, the docking score of these molecules can be calculated using a deep learning-based predictive model. The final aim was to combine both the target-specific generative model and the predictive model using reinforcement learning, so that the combined model can design molecules with an optimized docking score. A docking score predictive model was chosen, as it is several orders of magnitude faster than the actual docking process.[43,44] After an extensive hyperparameter tuning, the predictive model could predict the docking score for SMILES strings, within a root-mean-square error (rmse) of 0.5 (a comparison between the predicted and observed docking scores is provided at the end of this section, which explains the effectiveness of the predictive model).

Next, the target-specific generative model obtained after transfer learning was subjected to docking score optimization with the predictive model using regularized reinforcement learning (step d, Figure 1) for 70 epochs. The distribution of predicted docking scores before and after reinforcement learning was considered as the criterion for terminating the training process (Figure 2b). After reinforcement learning, 10,000 molecules were sampled from the trained generative model; 93% (9290) of the generated small molecules were found to be chemically valid (a SMILES is considered to be chemically valid if it can be parsed by the RDKit library[17]). Upon the removal of redundant (15.76%) and training set-identical molecules (2.45%), 7469 small molecules were obtained.

Next, various drug-like physicochemical property filters were applied to obtain molecules with drug-like properties. With the application of property filters (200 Da < molecular weight <700 Da, octanol−water partition coefficient <6.0, and synthetic accessibility score <5.0), a dataset of 6691 molecules was obtained. Synthetic accessibility score was used as a filter to avoid molecules which are difficult to be synthesized. Rule-based filters were applied to remove potentially problematic substructural features among the generated small molecules; 4.5% (455) molecules were found to be flagged by at least two of the four rule-based filters (PAINS, BRENK, NIH, and ZINC). The most common structural flags from each filter are provided in Supporting Information 1 (Table S2). Out of the remaining 6236 small molecules, 6106 small molecules with a docking score ≤ −7.0 were considered for further analysis.

While the distribution of the quantitative estimate of drug-likeness score (Figure 2c) remains similar, the final set of molecules was also found to have better synthetic accessibility scores[29] than the training dataset (Figure 2d). All the physicochemical property distributions of the generated small molecules in comparison with the training dataset are provided in Supporting Information 1 (Figure S3). It should also be noted that despite being not explicitly optimized, the generated molecules show physicochemical property distributions similar to that of the training dataset (Figure 2c,d).

Only 1.2% of the molecules generated after reinforcement learning were removed from the final list for poor docking score, indicating the effectiveness of reinforcement learning in optimizing the generative model. A Pearson correlation coefficient (r) of 0.67 was observed between the predicted scores and scores from AutoDock Vina,[21] indicating a high positive correlation between the predicted and observed values (Figure 3).

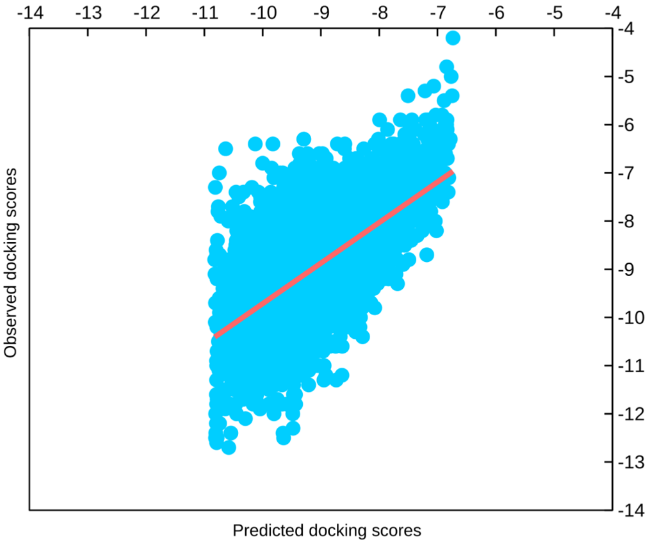


Figure 3. Comparison between predicted and observed docking scores: a linear regression-based comparison to illustrate the agreement between the model predicted and observed docking scores specific to the JAK2 active site. The red line indicates the regression line fitted for the input dataset of predicted and observed docking scores.

Generative Model Captures the Features of the Training Dataset. To understand the substructural features of the generated molecules, the fragment library in RDKit[17] was used. The average frequency of various fragments in the molecules generated by the pretrained generative model, transfer learning model, and the reinforcement learning model was computed by sampling 10,000 molecules in 10 batches of 1000 molecules each. They were compared with the training dataset, validation dataset, and the ChEMBL dataset. As the training and the validation dataset consisted of less than 10,000 molecules, random sampling with replacement was used to augment the dataset for the calculation. Frequencies of different fragments were calculated for each batch, and the average fragment frequency for all the batches was calculated. The average frequency of the top 10 fragments is tabulated in Table 1. Interestingly, all the top 10 fragments identified from our analysis are commonly utilized in the design and synthesis of highly selective JAK2 inhibitors.[45−51] Notably, tertiary amines have shown to increase the selectivity and ease of synthesis of JAK2 inhibitors,[45] and bicyclic groups are known to increase the selectivity of inhibitors toward JAK2 over JAK1, JAK3, and TYK2.[51]

The average fragment frequencies after transfer learning and reinforcement learning were compared with the training dataset and validation dataset frequencies. The average fragment frequencies from the model after transfer learning resemble that of the training dataset, whereas the average fragment frequencies from the model after reinforcement learning resemble that of the validation dataset. For instance, the frequency of tertiary amines after transfer learning was 3679, whereas the frequency after reinforcement learning was 3216. The decrease in frequency after reinforcement learning is an

Table 1. Model After Reinforcement Learning and the Pretrained Generative Model Learnt the Inherent Grammar of Their Respective Training Datasets: the Average Frequency of Top 10 Fragments in the Validation Dataset (JAK2 Inhibitors) and Molecules from the Generative Model Obtained After Reinforcement Learning (RL)[a]

S. no	Fragments	ChEMBL dataset	pretrained generative model	training dataset	model after TL	model after RL (final model)	validation dataset
1	tertiary amines	2046	1914.2	4141	3679	3216	2893.9
2	aromatic nitrogens	1345.3	1280.1	3269.7	2742.4	2513.5	2540
3	anilines	634.7	565.9	1549.2	1383.2	1275.3	1418.8
4	bicyclic groups	882.2	807.2	1251.4	979.8	1322.2	1391.3
5	secondary amines	891.6	831.6	1451.7	1286.5	1346	1331.5
6	benzene rings	1516.6	1436.7	868.4	942.5	1485.5	1094.8
7	halogens	702.4	704.7	784.6	849.9	826	604.6
8	carbonyl oxygens	1058.8	1063.3	897.3	906.1	668.4	592.8
9	carbonyl oxygen without COOH	966.9	961.9	892.8	892.7	649.3	577.3
10	amides	779.8	754.2	863.4	865.8	675.3	572.8

[a] The corresponding fragment frequencies from the ChEMBL dataset, the pretrained generative model, the transfer learning training dataset (JAK1, JAK3, and TYK2 inhibitors), and the model after transfer learning (TL) are also shown.

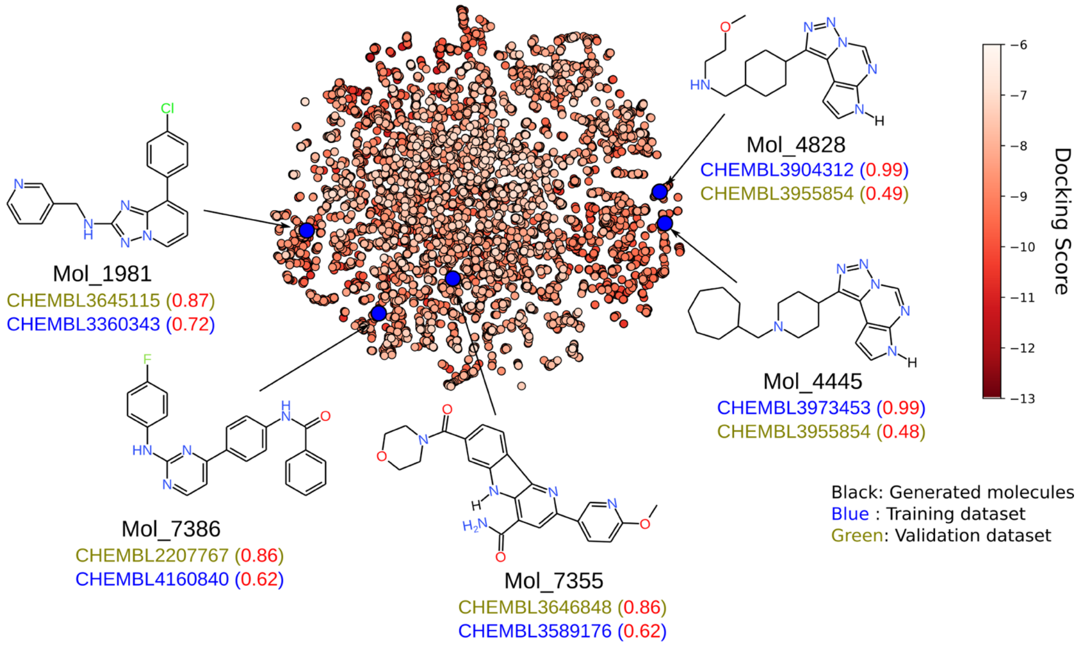


Figure 4. Optimized generative model learns from the training dataset and generates molecules close to the validation dataset: a low-dimensional embedding of the final set of generated small molecules colored based on their docking score. The five small molecules generated by the model (black) are shown according to their similarity toward either the training set (blue) or the validation set (green). The ChEMBL ID of the small molecules from the training and validation datasets is also shown. The Tanimoto coefficient of the generated small molecules with the training and validation dataset molecules is shown in red within parenthesis.

indication that the model is generating molecules closer to the validation dataset rather than the training dataset because of the reinforcement learning optimization. The small molecules from the reinforcement learning model that passed various filters are shown as a low-dimensional embedding (Figure 4). From the figure, it can be inferred that the generative model was able to produce small molecules with high similarity (quantified using the Tanimoto coefficient) to both the training dataset and the validation dataset. Molecules with both high similarity to the validation dataset and lower similarity to the training dataset were found among the generated small molecules. This also emphasizes that the optimized generative model samples from a region of the chemical space closer to the validation dataset.

Similarity of the Generated Small Molecules to the Validation Dataset. To further understand the distribution of the generated molecules with respect to the validation dataset, a lower dimensional embedding[52] is shown (Figure 5a). Extended Connectivity FingerPrint (ECFP4) was used as the descriptor for the embedding. From the embedding, it can be inferred that although being similar to a subset of the validation dataset, the molecules have also been sampled from a more optimized subspace of the chemical space occupied by the unique JAK2 inhibitors (Figure 5a). This can be substantiated by the ability of the generative model to generate molecules with higher-than-expected docking scores and with better physicochemical properties post reinforcement learning.

The final set of 6106 molecules obtained after virtual screening was compared against the validation dataset of unique JAK2 inhibitors. It was observed that 310 molecules (5%) from the generated set of molecules have the Tanimoto coefficient[23]

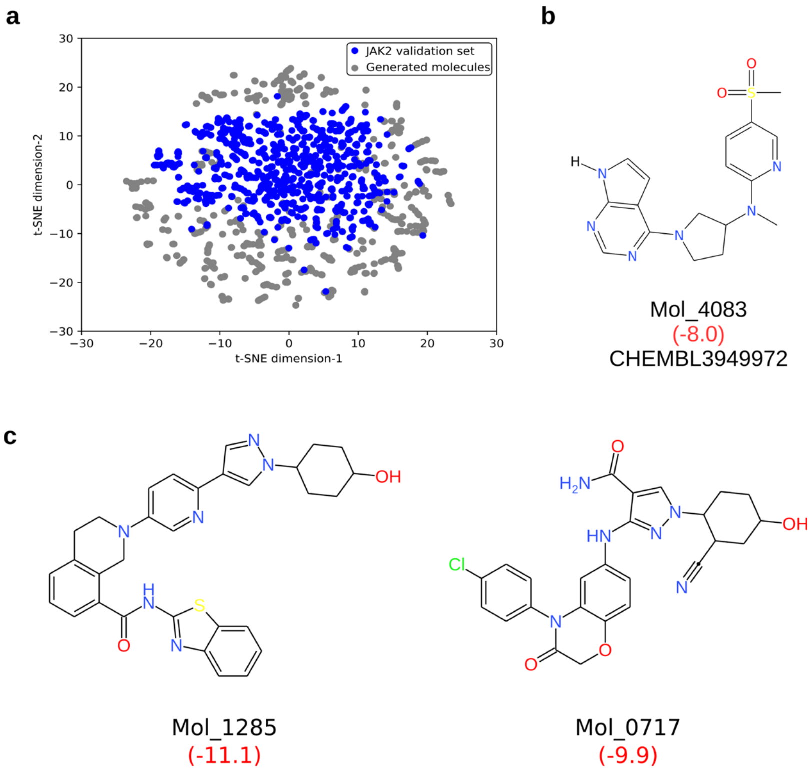


Figure 5. Generated small molecules identical to the validation dataset and molecules with improved docking scores: (a) *t*-distributed stochastic neighbor embedding (*t*-SNE) of the molecules generated by the generative model after reinforcement learning and the JAK2 validation dataset. (b) Small molecules generated by the generative model found to be identical to an existing JAK2 inhibitor. (c) Small molecules from the generative model with higher virtual screening scores compared to the molecules from the JAK2 validation dataset. The virtual screening scores from AutoDock Vina[21] are shown in red within parentheses.

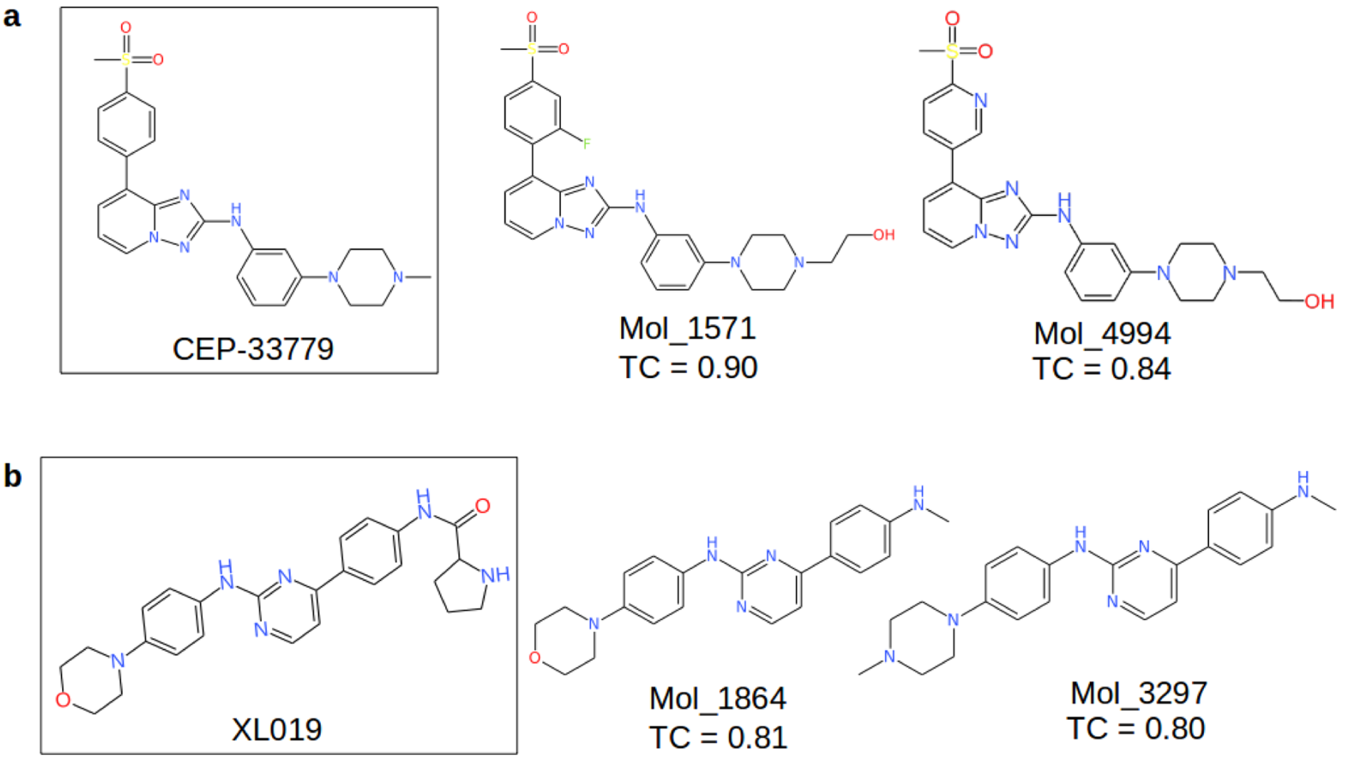


Figure 6. Subset of the generated small molecules with a high similarity to the existing JAK2-selective inhibitors (CEP-33779 and XL019). The similarity between the molecules has been quantified with the Tanimoto coefficient (TC).

above 0.75, relative to the molecules from the validation dataset, indicating high similarity. Out of these 310 molecules, one molecule was also found to be identical to the validation dataset (Figure 5b). Based on the virtual screening scores, it was observed that the new molecules can be potentially better inhibitors of JAK2 compared to the known inhibitor molecules. A subset of these representative molecules is shown in Figure 5c, and all the filtered molecules with their virtual screening scores are provided in Supporting Information 2.

Although the docking score can be used as a preliminary filter, the molecular mechanics Poisson−Boltzmann surface area (MM/PBSA) energies are more accurate when compared to any of the docking scoring schemes.[53] The MM/PBSA calculations were performed on all the three molecules (in Figure 5b,c). According to the MM/PBSA calculations, the predicted binding energy for the new molecules (−87.6 and −75.3 kJ/mol for Mol_1285 and Mol_0717, respectively) was better when compared to the existing JAK2 inhibitor (−28.9 kJ/mol). Based on the docking score and MM/PBSA calculations, it is evident that the new molecules designed using our pipeline are better binders when compared to the existing JAK2 inhibitor (CHEMBL3949972).

■ DISCUSSION

The aim of the current study is to generate small molecules with optimized docking score and desired properties for a target protein where no experimentally verified ligand dataset is available. As no bioactivity data exist for the ligand dataset, it is challenging to comment on the selectivity of the designed molecules. Nevertheless, to check whether the designed molecules have similarity with the JAK2-selective inhibitors, the existing literature was searched to identify the molecular features that are selective to JAK2.[45−51] Some of these features include the presence of aminothiazole, aminopyrazole, triazolo pyridine, or diaminopyrimidine in the head group. The presence of 4-fluorobenzyl, dicyclopropyl, carboxamide, morpholine, and cyclohexanol subgroups has been reported to enhance JAK2 selectivity.[48,49] Several generated small molecules were identified with the above features. A representative set of small molecules with a high docking score is shown in Supporting Information Figure S4 (see Supporting Information 1—Section S3 for a more detailed description).

It was observed that a set of generated small molecules is similar to at least two highly selective JAK2 inhibitors, CEP-33779[54] (Figure 6a) and XL019[55] (Figure 6b). CEP-33779 is a selective JAK2 inhibitor with an IC_{50} of 1.8 nM and >40- and >800-fold selectivity over JAK1 and TYK2, respectively.[49,54] XL019 is another selective JAK2 inhibitor, with IC_{50} of 2.2, 134.3, and 214.2 nM for JAK2, JAK1, and JAK3, respectively.[55] The designed molecules that are similar to CEP-33779 and XL019 are shown in Figure 6.

A diversity analysis was performed to understand if the model after reinforcement learning has generated any novel JAK2 scaffolds. The unique Murcko scaffolds[56] present in the training dataset, validation dataset, and the filtered set of generated molecules were extracted and compared using the Tanimoto coefficient (see Supporting Information 1—Section S4 for more details). It was observed that 25.21% of the scaffolds from the generated molecules were novel when compared to both the training and validation datasets. A representative set of novel scaffolds obtained using Butina clustering[57] is shown in Supporting Information 1 (Figure S5).

Advantages of Using the Proposed Deep Learning-Based Approach. There are three primary advantages of the proposed deep learning approach in this work.

Efficient Exploration of the Chemical Space. Traditional de novo drug design approaches mainly focus on the generation of novel small molecules with a high scaffold similarity to the existing inhibitors.[2] However, deep learning models have shown the ability to generate completely novel scaffolds and small molecules through the generative model. This is evident from the observation that 25.21% of the scaffolds from the generated molecules are novel when compared to both the training and validation datasets.

Target-Specific Molecule Design. By using transfer learning, deep learning models are capable of capturing the pharmacophoric representations from a dataset of target-specific small molecules.

Dynamic Control over PhysicoChemical Properties. It is extremely difficult to modulate the properties of small molecules while designing them using traditional approaches. Although, in this particular work, docking score was used for the predictive model, it is possible to replace the docking score with the desired physicochemical properties for on-the-fly property optimization using reinforcement learning.

In order to understand the effectiveness of using both transfer and reinforcement learning to train the generative model, additional tests were performed using only transfer learning and only reinforcement learning, respectively (see Supporting Information 1—Section S5 for details). During only transfer learning, sufficient shift in the distribution of the Tanimoto coefficient was observed, but there was no improvement in the docking score (Figure S6). In contrast, during only reinforcement learning, an improvement in the docking score distribution was observed, but there was no shift in the Tanimoto coefficient distribution (Figure S6). However, when both transfer and reinforcement learning were used together, an improvement in both the docking score and Tanimoto coefficient distribution could be observed (Figure 2a,b).

■ CONCLUSIONS

We have proposed a *de novo* drug design method for generating small molecules against targets of interest where no target-specific small-molecule dataset is available. The proposed method is useful for target proteins whose structure is available/modeled and the active site is known. The sequence information of the target protein has been used to identify the related proteins whose inhibitors are used as the target-specific training dataset. Once these two prerequisites are satisfied, the proposed method can be applied to any target protein. Curating a diverse target-specific ligand dataset enables the generative model to learn all the necessary molecular features required to bind to the active site of the target protein. A docking-based predictive model was trained to optimize the docking score of the generated molecules. Post optimization, the generated small molecules were subjected to various physicochemical property filters and rule-based filters to remove potentially problematic compounds with undesired properties. The proposed approach can be extended to design molecules against proteins with a conserved function, for example, in the case of virus or pathogens.

It was observed that the target-specific generative model was able to translate the features learnt from the training dataset to generate molecules similar to the validation dataset. By utilizing both transfer learning and reinforcement learning in sequence, the affinity of the designed small molecules toward the JAK2 active site was optimized. The effectiveness of optimization using reinforcement learning is visible from the generated small molecules with improved virtual screening scores compared to the validation dataset. The analysis of the substructural fragments present among the generated small molecules also indicates that the molecules are selective against the JAK2 active site. Although the method was validated by its ability to reproduce molecules from the validation dataset, several molecules with better predicted binding capability (quantified by their docking score and MM/PBSA energies) were also observed.

■ ASSOCIATED CONTENT

SI Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.0c01060.

> Supporting Information 1: details of the hyperparameters used to train various deep learning models, comparison of Janus kinase family of proteins, effectiveness of using both transfer and reinforcement learning, generated novel scaffolds, representative set of small molecules, and various physicochemical property distributions of the generative models (PDF)
>
> Supporting Information 2: training and validation datasets for the case study and the dataset of generated small molecules obtained after application of all filters along with their virtual screening scores (XLS)

■ AUTHOR INFORMATION

Corresponding Authors

Gopalakrishnan Bulusu − *TCS Innovation Labs-Hyderabad (Life Sciences Division), Tata Consultancy Services Limited, Hyderabad 500081, India;* orcid.org/0000-0002-4958-7573; Email: g.bulusu@tcs.com

Arijit Roy − *TCS Innovation Labs-Hyderabad (Life Sciences Division), Tata Consultancy Services Limited, Hyderabad 500081, India;* orcid.org/0000-0002-1961-2483; Email: roy.arijit3@tcs.com

Authors

Sowmya Ramaswamy Krishnan − *TCS Innovation Labs-Hyderabad (Life Sciences Division), Tata Consultancy Services Limited, Hyderabad 500081, India;* orcid.org/0000-0001-5404-3266

Navneet Bung − *TCS Innovation Labs-Hyderabad (Life Sciences Division), Tata Consultancy Services Limited, Hyderabad 500081, India;* orcid.org/0000-0002-6376-277X

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.0c01060

Author Contributions

[†]S.R.K. and N.B. with equal contribution.

Notes

The authors declare the following competing financial interest(s): All the authors are employed at Tata Consultancy Services Ltd.

The code used to generate results shown in this study is available from the corresponding author for academic use.

■ ACKNOWLEDGMENTS

We thank our colleagues Dr. Rajgopal Srinivasan, Dr. Siladitya Padhi, Dibyajyoti Das, and Dr. Broto Chakrabarty for helpful discussions.

■ REFERENCES

(1) Chen, J.; Luo, X.; Qiu, H.; Mackey, V.; Sun, L.; Ouyang, X. Drug Discovery and Drug Marketing with the Critical Roles of Modern Administration. *Am. J. Transl. Res.* **2018**, *10*, 4302−4312.

(2) Stahl, M.; Todorov, N. P.; James, T.; Mauser, H.; Boehm, H.-J.; Dean, P. M. A Validation Study on the Practical Use of Automated De Novo Design. *J. Comput. Aided Mol. Des.* **2002**, *16*, 459−478.

(3) Walters, W. P. Virtual Chemical Libraries: Miniperspective. *J. Med. Chem.* **2018**, *62*, 1116−1124.

(4) Colling, R.; Pitman, H.; Oien, K.; Rajpoot, N.; Macklin, P.; Snead, D.; Sackville, T.; Verrill, C.; CM-Path AI in Histopathology Working Group. A Roadmap to Routine Use in Clinical Practice. *J. Pathol.* **2019**, *249*, 143−150.

(5) Ho, C. W. L.; Soon, D.; Caals, K.; Kapur, J. Governance of Automated Image Analysis and Artificial Intelligence Analytics in Healthcare. *Clin. Radiol.* **2019**, *74*, 329−337.

(6) Yu, K.-H.; Beam, A. L.; Kohane, I. S. Artificial Intelligence in Healthcare. *Nat. Biomed. Eng.* **2018**, *2*, 719−731.

(7) Mak, K.-K.; Pichika, M. R. Artificial Intelligence in Drug Development: Present Status and Future Prospects. *Drug Discov. Today* **2019**, *24*, 773−780.

(8) Segler, M. H. S.; Kogej, T.; Tyrchan, C.; Waller, M. P. Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks. *ACS Cent. Sci.* **2017**, *4*, 120−131.

(9) Olivecrona, M.; Blaschke, T.; Engkvist, O.; Chen, H. Molecular De-Novo Design Through Deep Reinforcement Learning. *J. Cheminf.* **2017**, *9*, 48.

(10) Popova, M.; Isayev, O.; Tropsha, A. Deep Reinforcement Learning for De Novo Drug Design. *Sci. Adv.* **2018**, *4*, No. eaap7885.

(11) Born, J.; Manica, M.; Oskooei, A.; Cadow, J.; Martínez, M. R. Paccmann^{RL}: Designing Anticancer Drugs from Transcriptomic Data via Reinforcement Learning. *Proceedings of the International Conference on Research in Computational Molecular Biology*, June 22−25, 2020; pp 231−233.

(12) Li, Y.; Zhang, L.; Liu, Z. Multi-Objective De Novo Drug Design with Conditional Graph Generative Model. *J. Cheminf.* **2018**, *10*, 33.

(13) Ståhl, N.; Falkman, G.; Karlsson, A.; Mathiason, G.; Boström, J. Deep Reinforcement Learning for Multiparameter Optimization in De Novo Drug Design. *J. Chem. Inf. Model.* **2019**, *59*, 3166−3176.

(14) Joulin, A.; Mikolov, T. Inferring Algorithmic Patterns with Stack-Augmented Recurrent Nets. *Proceedings of the International Conference on Neural Information Processing Systems*: Montréal, Canada, December 7−12, 2015; Vol. *1*; pp 190−198.

(15) *ChEMBL Database*, version 27, 10.6019/CHEMBL.database.27 (accessed November, 2019).

(16) Weininger, D.; Weininger, A.; Weininger, J. L. SMILES. 2. Algorithm for Generation of Unique SMILES Notation. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 97−101.

(17) Landrum, G. *RDKit: Open-Source Cheminformatics Software*. http://www.rdkit.org (accessed Nov 2019).

(18) Chung, J.; Gulcehre, C.; Cho, K.; Bengio, Y. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *NIPS 2014 Workshop on Deep Learning*: Montréal, Canada, December 8−13, 2014.

(19) Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. Basic Local Alignment Search Tool. *J. Mol. Biol.* **1990**, *215*, 403−410.

(20) Pearson, W. R. An Introduction to Sequence Similarity ("homology") Searching. *Curr. Protoc. Bioinf.* **2013**, *42*, 3.1.1.

(21) Trott, O.; Olson, A. J. AutoDock Vina: Improving the Speed and Accuracy of Docking with A New Scoring Function, Efficient Optimization and Multithreading. *J. Comput. Chem.* **2010**, *31*, 455−461.

(22) Tan, C.; Sun, F.; Kong, T.; Zhang, W.; Yang, C.; Liu, C. A Survey on Deep Transfer Learning. *Proceedings of the International Conference on Artificial Neural Networks*: Rhodes, Greece, October 4−7, 2018; pp 270−279.

(23) Lipkus, A. H. A Proof of the Triangle Inequality for the Tanimoto Distance. *J. Math. Chem.* **1999**, *26*, 263−265.

(24) Kingma, D. P.; Ba, J. Adam: A Method for Stochastic Optimization. *Proceedings of International Conference on Learning Representations*: San Diego, California, May 7−9, 2015.

(25) Sutton, R. S.; Barto, A. G. *Reinforcement Learning: An Introduction*, 1st ed.; MIT Press: U.S., 1998.

(26) Goodfellow, I. J.; Mirza, M.; Xiao, D.; Courville, A.; Bengio, Y. An Empirical Investigation of Catastrophic Forgetting in Gradient-Based Neural Networks. *Proceedings of International Conference on Learning Representations*: Banff, Alberta, Canada, April 14−16, 2014.

(27) Jaques, N.; Gu, S.; Turner, R. E.; Eck, D. Tuning Recurrent Neural Networks with Reinforcement Learning. *Proceedings of International Conference on Learning Representations*: Toulon, France, April 24−26, 2017.

(28) Tran, P. T.; Phong, L. T. On the Convergence Proof of AMSGrad and a New Version. *IEEE Access* **2019**, *7*, 61706−61716.

(29) Ertl, P.; Schuffenhauer, A. Estimation of Synthetic Accessibility Score of Drug-Like Molecules Based on Molecular Complexity and Fragment Contributions. *J. Cheminf.* **2009**, *1*, 8.

(30) Baell, J. B.; Holloway, G. A. New Substructure Filters for Removal of Pan Assay Interference Compounds (PAINS) from Screening Libraries and for their Exclusion in Bioassays. *J. Med. Chem.* **2010**, *53*, 2719−2740.

(31) Brenk, R.; Schipani, A.; James, D.; Krasowski, A.; Gilbert, I. H.; Frearson, J.; Wyatt, P. G. Lessons Learnt from Assembling Screening Libraries for Drug Discovery for Neglected Diseases. *ChemMedChem* **2008**, *3*, 435.

(32) Doveston, R. G.; Tosatti, P.; Dow, M.; Foley, D. J.; Li, H. Y.; Campbell, A. J.; House, D.; Churcher, I.; Marsden, S. P.; Nelson, A. A Unified Lead-Oriented Synthesis of over Fifty Molecular Scaffolds. *Org. Biomol. Chem.* **2015**, *13*, 859−865.

(33) van Der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. GROMACS: Fast, Flexible, and Free. *J. Comput. Chem.* **2005**, *26*, 1701−1718.

(34) Huang, J.; MacKerell, A. D., Jr. CHARMM36 All-Atom Additive Protein Force Field: Validation Based on Comparison to NMR Data. *J. Comput. Chem.* **2013**, *34*, 2135−2145.

(35) Vanommeslaeghe, K.; Raman, E. P.; MacKerell, A. D. Automation of the CHARMM General Force Field (CGenFF) II: Assignment of Bonded Parameters and Partial Atomic Charges. *J. Chem. Inf. Model.* **2012**, *52*, 3155−3168.

(36) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **1983**, *79*, 926.

(37) Bussi, G.; Donadio, D.; Parrinello, M. Canonical Sampling through Velocity Rescaling. *J. Chem. Phys.* **2007**, *126*, 014101.

(38) Kumari, R.; Kumar, R.; Lynn, A.; Open Source Drug Discovery Consortium. G_mmpbsa–a GROMACS Tool for High-Throughput MM-PBSA Calculations. *J. Chem. Inf. Model.* **2014**, *54*, 1951−1962.

(39) Yamaoka, K.; Saharinen, P.; Pesu, M.; Holt, V. E. T.; Silvennoinen, O.; O'Shea, J. J. The Janus Kinases (Jaks). *Genome Biol.* **2004**, *5*, 253.

(40) Rawlings, J. S.; Rosler, K. M.; Harrison, D. A. The JAK/STAT Signaling Pathway. *J. Cell Sci.* **2004**, *117*, 1281−1283.

(41) Vainchenker, W.; Constantinescu, S. N. JAK/STAT Signaling in Hematological Malignancies. *Oncogene* **2013**, *32*, 2601−2613.

(42) UniProt Consortium. UniProt: A Worldwide Hub of Protein Knowledge. *Nucleic Acids Res.* **2019**, *47*, D506−D515.

(43) Jastrzębski, S.; Szymczak, M.; Pocha, A.; Mordalski, S.; Tabor, J.; Bojarski, A. J.; Podlewska, S. Emulating Docking Results Using a Deep Neural Network: A New Perspective for Virtual Screening. *J. Chem. Inf. Model.* **2020**, *60*, 4246−4262.

(44) Gentile, F.; Agrawal, V.; Hsing, M.; Ton, A.-T.; Ban, F.; Norinder, U.; Gleave, M. E.; Cherkasov, A. Deep Docking: A Deep Learning Platform for Augmentation of Structure Based Drug Discovery. *ACS Cent. Sci.* **2020**, *6*, 939−949.

(45) Douglas, J. J.; Cole, K. P.; Stephenson, C. R. J. Photoredox Catalysis in a Complex Pharmaceutical Setting: Toward the Preparation of JAK2 Inhibitor LY2784544. *J. Org. Chem.* **2014**, *79*, 11631−11643.

(46) Lin, T. E.; HuangFu, W.; Chao, M.; Sung, T.; Chang, C.; Chen, Y.; Hsieh, J.; Tu, H.; Huang, H.; Pan, S.; Hsu, K. A Novel Selective JAK2 Inhibitor Identified Using Pharmacological Interactions. *Front. Pharmacol.* **2018**, *9*, 1379.

(47) Hanan, E. J.; van Abbema, A.; Barrett, K.; Blair, W. S.; Blaney, J.; Chang, C.; Eigenbrot, C.; Flynn, S.; Gibbons, P.; Hurley, C. A.; Kenny, J. R.; Kulagowski, J.; Lee, L.; Magnuson, S. R.; Morris, C.; Murray, J.; Pastor, R. M.; Rawson, T.; Siu, M.; Ultsch, M.; Zhou, A.; Sampath, D.; Lyssikatos, J. P. Discovery of Potent and Selective Pyrazolopyrimidine Janus Kinase 2 Inhibitors. *J. Med. Chem.* **2012**, *55*, 10090−10107.

(48) Dymock, B. W.; See, C. S. Inhibitors of JAK2 and JAK3: An Update on the Patent Literature 2010−2012. *Expert Opin. Ther. Pat.* **2013**, *23*, 449−501.

(49) Dymock, B. W.; Yang, E. G.; Chu-Farseeva, Y.; Yao, L. Selective JAK Inhibitors. *Future Med. Chem.* **2014**, *6*, 1439−1471.

(50) Su, Q.; Ioannidis, S.; Chuaqui, C.; Almeida, L.; Alimzhanov, M.; Bebernitz, G.; Bell, K.; Block, M.; Howard, T.; Huang, S.; Huszar, D.; Read, J. A.; Costa, C. R.; Shi, J.; Su, M.; Ye, M.; Zinda, M. Discovery of 1-methyl-1H-imidazole Derivatives as Potent Jak2 Inhibitors. *J. Med. Chem.* **2014**, *57*, 144−158.

(51) Ma, L.; Clayton, J. R.; Walgren, R. A.; Zhao, B.; Evans, R. J.; Smith, M. C.; Heinz-Taheny, K. M.; Kreklau, E. L.; Bloem, L.; Pitou, C.; Shen, W.; Strelow, J. M.; Halstead, C.; Rempala, M. E.; Parthasarathy, S.; Gillig, J. R.; Heinz, L. J.; Pei, H.; Wang, Y.; Stancato, L. F.; Dowless, M. S.; Iversen, P. W.; Burkholder, T. P. Discovery and Characterization of LY2784544, a Small-Molecule Tyrosine Kinase Inhibitor of JAK2V617F. *Blood Canc. J.* **2013**, 3, No. e109.

(52) van der Maaten, L.; Hinton, G. Visualizing Data Using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579−2605.

(53) Wang, E.; Sun, H.; Wang, J.; Wang, Z.; Liu, H.; Zhang, J. Z. H.; Hou, T. End-point Binding Free Energy Calculation with MM/PBSA and MM/GBSA: Strategies and Applications in Drug Design. *Chem. Rev.* **2019**, *119*, 9478−9508.

(54) Dugan, B. J.; Gingrich, D. E.; Mesaros, E. F.; Milkiewicz, K. L.; Curry, M. A.; Zulli, A. L.; Dobrzanski, P.; Serdikoff, C.; Jan, M.; Angeles, T. S.; Albom, M. S.; Mason, J. L.; Aimone, L. D.; Meyer, S. L.; Huang, Z.; Wells-Knecht, K. J.; Ator, M. A.; Ruggeri, B. A.; Dorsey, B. D. A Selective, Orally Bioavailable 1, 2, 4-triazolo [1, 5-a] pyridine-based Inhibitor of Janus kinase 2 for Use in Anticancer Therapy: Discovery of CEP-33779. *J. Med. Chem.* **2012**, *55*, 5243−5254.

(55) Forsyth, T.; Kearney, P. C.; Kim, B. G.; Johnson, H. W. B.; Aay, N.; Arcalas, A.; Brown, D. S.; Chan, V.; Chen, J.; Du, H.; Epshteyn, S.; Galan, A. A.; Huynh, T. P.; Ibrahim, M. A.; Kane, B.; Koltun, E. S.; Mann, G.; Meyr, L. E.; Lee, M. S.; Lewis, G. L.; Noguchi, R. T.; Pack, M.; Ridgway, B. H.; Shi, X.; Takeuchi, C. S.; Zu, P.; Leahy, J. W.; Nuss, J. M.; Aoyama, R.; Engst, S.; Gendreau, S. B.; Kassees, R.; Li, J.; Lin, S.-H.; Martini, J.-F.; Stout, T.; Tong, P.; Woolfrey, J.; Zhang, W.; Yu, P. SAR and in vivo Evaluation of 4-aryl-2-aminoalkylpyrimidines as Potent and Selective Janus Kinase 2 (JAK2) Inhibitors. *Bioorg. Med. Chem. Lett.* **2012**, *22*, 7653−7658.

(56) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887−2893.

(57) Butina, D. Unsupervised Data Base Clustering Based on Daylight's Fingerprint and Tanimoto Similarity: A Fast and Automated Way To Cluster Small and Large Data Sets. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 747−750.